**Data Storytelling**

**~ Advait Menon, J045**

**Chosen Dataset: Drug Poisoning Mortality by State**

**Elaborating the data set and its attributes**

The chosen data contains two layers of data. One is the overall state-wide drug poisoning mortality rate, the other one being the mortality rate for specific racial segments of the United States society (viz. Hispanic, Hispanic Black/White etc). The data attributes include Year (to show yearly statistics), Sex (to differentiate between the statistics of different genders within the various racial segments), Race and Hispanic Origin, State (includes all 50 states and DC (District of Columbia) as a separate region), Deaths (in number, in the corresponding year, corresponding state OR corresponding Race & Hispanic Origin), Population (of the state/race in that year), Crude Death Rate (Deaths dependent on population of a state/race), Standard Error for Crude Rate, Lower Confidence Limit for Crude Rate, Upper Confidence Limit for Crude Rate, Age-adjusted Rate, Standard Error, Age-adjusted Rate, Lower Confidence Limit for Age-adjusted rate, Upper Confidence Limit for Age-adjusted Rate, State Crude Rate in Range, US Crude Rate and US Age-adjusted Rate.

**Steps taken for cleaning of data**

For purposes of inclusivity and ease of handling of data, the data points segregating the citizens on the grounds of their race were isolated and excluded from the raw dataset, thereby leaving only the mortality rates based on the US states in each year. Moreover, for the sake of relevance in the story to be woven around the chosen dataset, the statistical attributes, *save* Deaths, Population, and Crude Death Rate, were removed. Following this, the data was chronologically sorted (using year as the sorting attribute), and different steps were taken to ensure that the dataset is consistent and clean. This included checks for unnecessary spaces in every attribute/column, spell checks to ensure consistency of names, continuity in, and range-appropriate, years (i.e., making sure the data only contains all its years between 1999-2015), checking for the inclusion of all 50 states of the US (plus District of Columbia, making it 51 data points per year) and ensuring there are no other empty data points.

Having done with the cleaning process, the data was then imported from the updated Excel file to Tableau for visualisation purposes.

**Visualisations chosen and their interpretations**

**Slide 1: Yearly Drug Poisoning Stats**

This is the most obvious visualisation that paints a picture of how drug poisoning has affected the United States population in each year. This graph has an increasing trend overall, with a sum total of 52404 total deaths by 2015. The Gantt Bar between the same attributes on the same slide shows a substantial difference between the death count in 2000-2001(11.36%), 20014-2015 (11.37%) and 2001-2002(21.26%), amongst other increases.

**Slide 2: Crude Death Rate vs Year**

The Crude Death Rate, being dependent on the population of the state/region being investigated, also shows an increasing rate despite decreasing in some year ranges. This increasing trend is, once again, proved using a Gantt Bar showing percent differences.

**Slide 3: Overall State-wise Drug Poisoning Statistics**

This is implemented using a horizontal (decreasing) bar chart that shows an overview of the drug poisoning cases in all the 50+1 regions of the United States over all of the years in the range 1999-2015. Considering that this bar chart is plotted between states and sum of death (in numbers) and *not* crude death rate, it removes the need of any dependency of the deaths on population density in the States, otherwise created by Crude Death Rate. In this way, it can be inferred that California is most affected by the crisis of drug poisoning (with a sum total of 59,427 deaths between 1999 and 2015). On the other hand, North Dakota is the least affected by the same (with a total of 411 deaths over the same range of years).

**Slide 4: Filled Map**

In the filled map on Slide 4, the darker shades of blue (such as California, Texas, Florida) represent higher mortality rates, while the lighter shades of blue (such as Montana, North Dakota, Wyoming) represent lower drug poisoning deaths. However, on further analysis of states and drug poisoning cases, by means of Crude Death Rate, it can be observed that the state with maximum death rates is West Virginia and not California. However, this may be attributed to the fact that West Virginia also has a lesser population than California but has more deaths within this relatively lesser population, thereby confirming that Crude Death Rate as a calculated measure is not necessarily a good measure of the real drug poisoning statistics.

**Slide 5: Boxplot of Deaths in states.**

The boxplot in slide 5 is plotted on the basis of the total deaths and their quartiles, while also indicating how different states stand in terms of the deaths in the years ranging from 1999 to 2015. It shows that in the discrete set of deaths in 50+1 territories, California, Florida and Texas are outliers with respect to the box-and-whiskers plot. On the other hand, Pennsylvania, New York and a few other states lie between the 2nd quartile (median) and 3rd quartile in terms of the deaths due to drug poisoning.

**Slide 6: A closer look into the state with the highest deaths over several years (California)**

The state of California is being analysed in slide 6, being the state with the largest drug poisoning-related deaths. As per the line graph and the adjacent Gantt graph both, it can be seen that California had a sharp decrease in its drug poisoning deaths in the range 1999-2000 and 2000-2001 (with Florida being the state having the highest cases of drug poisoning in 2001). However, the death count took a mammoth leap between 2001 and 2002 with a 142.7% increase in the deaths, which happens to be the largest increase in the entire Gantt Bar graph. Post this increase, the numbers have only increased, making California an outlier in drug poisoning cases as well.

**Summary**

In conclusion, it can be said that the case of drug poisoning is still on an all-time high and has only gone higher with each passing year.

The analysis conducted using the given data is just one method to analyse the trends in drug poisoning by means of utilising number of deaths as a parameter. With the availability of further data such as the variety/category/specific names of drugs that are being consumed, leading to these cases of drug poisoning, a further root cause analysis may be done as to why a certain drop or peak occurred (such as the peak in California in 2001-2002). The availability of this data along with other survey data such as the number of samples, mean etc. also means that further statistical methods such as hypothesis testing, and machine learning techniques like regression and classification may be conducted to not only analyse the trends, but also to prevent and inhibit the cases of drug poisoning using ample historical information that may be fed to an automated, supervised machine learning based program.